**The Use of Machine Learning Techniques to Identify and Deter Fraudulent Activities by Healthcare Providers**

A Project Report

Presented to

DATA-245-11

Spring, 2023

By

Bhavana Prasad Kote (016044899)

Jay Dattoo Dale (016646279)

Rashmi Shree Veeraiah (016099395)

Sawan Shivanand Beli (016522662)

May 18, 2023

**Table of contents**

Abstract

1. Introduction
   1. Problem Statement
   2. Project Background
   3. Literature Survey
2. CRISP-DM Approach
3. System Architecture
4. Data Preparation
   1. Data Exploration
   2. Data Preprocessing
   3. Data Transformation
   4. Data Normalization
5. Model Selection
   1. Models Implemented
   2. Training model and results
6. Deployment
7. Source Code Link

Conclusion

Future Scope

**Abstract**

Healthcare provider fraud is a major challenge in the medical field, resulting in monetary losses and impaired quality of care for patients. Fraudulent activities by healthcare providers such as unnecessary medical procedures, false diagnoses, and billing for services not rendered, result in the loss of billions of dollars annually.Improper payments, including fraudulent claims, accounted for an estimated $52 billion in Medicare spending in 2017, according to the US Government Accountability Office (GAO).This project employs machine learning (ML) algorithms such as Random Forest, Logistic Regression, Decision Trees, and XGBoost to identify possible fraudulent claims using a broad collection of historical healthcare-related data. Because of their shown success in classification tasks, Random Forest, Logistic Regression, Decision Trees, and XGBoost models are used. To assure data quality, the dataset is thoroughly preprocessed, and data transformation and feature engineering techniques such as normalization, label encoding, etc are used to extract useful information. The models are trained on the preprocessed dataset, which includes the demographics of patients, medical codes, billing information, and information about providers. To discriminate between fraudulent and non-fraudulent claims, the models learn patterns and relationships within the data. To analyze the performance of the models, evaluation criteria such as AUC, Accuracy, and F-1 score are used. The research intends to attain a high level of accuracy in recognizing possibly fraudulent claims by intensive experimentation and fine-tuning of the hyperparameters The knowledge gathered from this project can help to design powerful fraud detection systems in the healthcare business, allowing organizations to reduce financial losses, increase compliance, and improve patient care.

*Keywords*: Healthcare provider fraud, machine learning, random forest, logistic regression, decision trees, XGBoost, fraudulent claims detection

1. **Introduction**

There are voiced concerns about the rise in fraudulent activities, scams, and schemes perpetrated by doctors and hospitals in the recent past. The Department of Justice (DOJ) successfully collected more than 3 billion dollars through False Claims cases in 2019, with approximately 2.6 billion dollars related to healthcare fraud schemes. These healthcare fraud charges involved a variety of institutions, including medical instrument companies, service providers, hospitals, pharmacies, palliative organizations, laboratories, and physicians. The engagement of numerous parties emphasizes the problem's extensive depth and the necessity for greater attention and enforcement actions.

The objective of this project is to develop models that can predict potentially fraudulent healthcare providers based on the claims they file. The dataset used here is the historical healthcare-related data which comprises details of claims made by inpatients and outpatients, healthcare provider information, and the beneficiary details for each provider, sourced from Kaggle. Additionally, we aim to identify significant variables that aid in detecting the behavior of these potentially fraudulent providers. Furthermore, this study analyzes patterns in fraudulent claims submitted by providers to gain insights into their future behavior. Considering the prior data and research in identifying fraudulent behavior in the health industry, Decision Trees, Random Forest, Extreme Gradient Boosting, and Logistic Regression Machine Learning models are trained and tested. This research will contribute to a better understanding of fraudulent patterns, identify potentially fraudulent claims, and support the development of effective fraud detection systems in the healthcare industry.

* 1. ***Problem Statement***

Based on the selected health insurance plan and the policyholder's eligibility, health insurance firms have a crucial role in providing coverage for the medical costs of insurance holders. The scope of coverage varies according to the patient's diagnosis, course of treatment, and participating healthcare providers. Healthcare providers genuinely settle claims for their patients, there are instances where insured individuals or healthcare service providers attempt to defraud the system by submitting false claim details. These fraudulent practices can include fabricating bills, submitting duplicate bills, unnecessary treatments, and other deceptive actions, constituting a form of medical crime. As a result, insurance firms deliberately find up paying benefits for false claims, which not only results in loss of money but also makes it difficult to pay benefits for real claims that are true. The functioning of insurance firms are significantly impacted by these unethical actions, putting their capacity to properly assist policyholders and protect the integrity of the healthcare system.

* 1. ***Project Background***

This study utilizes the healthcare data procured for analyzing the fraud claims from Kaggle. It consists of inpatient and outpatient claims data that includes patient's diagnosis codes, procedure codes, claim amounts, etc., beneficiary details such as demographics, enrollment details, chronic conditions, etc of each healthcare provider, and details of providers submitting the claims. The task at hand involves a binary classification problem where the aim is to categorize data into two classes: fraudulent claims and legitimate claims in addition to identifying significant features that contribute to accurately predicting fraudulent claims. The project analyzes the claims data for the year 2018 performing exploratory data analysis, followed by data preprocessing and transformation, feature engineering, data normalization, and feature selection for the models that can help accurately classify claims as fraud or non-fraud.

Additionally, Machine Learning algorithms such as Logistic Regression, Random Forest, Decision Trees, and XGBOOST are trained on the preprocessed data and tested, followed by hyperparameter tuning. These models are evaluated using various evaluation metrics including AUC, F-1 score, and accuracy. Finally, the model with a lesser misclassification rate, highest AUC, and accuracy is selected as the most suitable and best model for classifying medical claims and identifying potentially fraudulent claims.

* 1. ***Literature Review***

Bauder and Khoshgoftaar (2017) present a comparative study of machine-learning methods for Medicare fraud detection. The study evaluates supervised, unsupervised, and hybrid approaches using four performance metrics and addresses class imbalance through oversampling and an 80-20 undersampling method. The 2015 Medicare data is grouped by provider types, and fraud labels are obtained from the List of Excluded Individuals/Entities database. The results demonstrate the successful detection of fraudulent providers, with the 80-20 sampling method showing the best performance across the learners. Supervised methods generally outperform unsupervised or hybrid methods, but the performance varies depending on the class imbalance sampling technique and provider type. The study introduces 10 specific techniques within the supervised, unsupervised, and hybrid learner groups, and evaluates their performance using different evaluation metrics. The supervised learners consist of Gradient Boosted Machine (GBM), Random Forest (RF), Deep Neural Network (DNN), and Naive Bayes (NB) models. The unsupervised methods include Autoencoder, Mahalanobis distance, k-Nearest Neighbors (kNN), and Local Outlier Factor (LOF). The hybrid learner's group combines a pre-trained neural network model using the unsupervised autoencoder technique, and another method that combines multivariate regression and Bayesian probability. The findings show that supervised learners perform better overall, but pre-training a model using an unsupervised autoencoder shows promise in achieving comparable performance with limited labeled data. The hybrid method also performs better than unsupervised methods.

The research paper by Johnson and Khoshgoftaar (2022) addresses the issue of healthcare fraud, which has significant financial and quality-of-care implications. The complexity and scale of healthcare systems make manual intervention in detecting fraud impractical. However, the abundance of electronic health record data provides an opportunity for data mining and machine learning approaches. Their study focuses on publicly available Medicare data, specifically the Medicare Physician and Other Practitioners data sets, which contain provider-level, claims-level, and beneficiary-level information. Two new labeled data sets are created for fraud classification, leveraging the provider summary data. The performance of these data sets is compared to an aggregated data set commonly used in related works. The evaluation involves using ensemble learners such as Random Forest and Extreme Gradient Boosting, and performance metrics include AUC, AUPRC, TPR, and TNR. The models are trained on a labeled dataset and evaluated using AUC and the results demonstrate that Gradient Boosting outperforms the other models with an AUC of 0.89, while Random Forests come in a close second with an AUC of 0.87. Statistical analysis confirms that the proposed data set outperforms the baseline. Additionally, the study identifies the top 20 features contributing to fraud prediction using SHAP values. The results demonstrate the effectiveness of the provider summary data in detecting Medicare fraud and highlight the superiority of the newly introduced data set.

Wu. Y et al. (2018) in their paper review over 70 studies from 2000 to 2017 and finds that the most commonly used machine learning models for healthcare fraud detection are Decision Trees, Artificial Neural Networks, Logistic Regression, Support Vector Machines, Naive Bayes Classifier, Random Forest, and K-Nearest Neighbors. The National Health Care Anti-Fraud Association (NHCAA) dataset, which contains healthcare claims data for suspected fraudulent activities, is used to train the models. They achieved an AUC (Area Under the Curve) score of 0.94 and an F1 score of 0.59, indicating a high level of accuracy in detecting fraudulent healthcare claims. The authors also found that the random forest model outperformed other machine learning models such as decision trees, logistic regression, and support vector machines. Limitations including the lack of own empirical results, leading to the need for further research to validate findings in real-world settings, and challenges such as data quality, class imbalance, and real-time detection are mentioned. The paper also emphasizes the need for using ensemble models, deep learning, and explainable AI techniques to address the challenges and improve the effectiveness of healthcare fraud detection systems.

Bauder et al. (2018b) proposed a novel unsupervised approach for identifying Medicare provider fraud using a publicly available dataset of Medicare claims data. A combination of clustering and anomaly detection techniques, K-Means clustering, and Local Outlier Factor (LOF) anomaly detection were used to implement the models for identifying the Medicare provider fraud. The proposed approach achieved a precision of 0.88, a recall of 0.35, and an F1 score of 0.50. The results illustrate that the clustering technique is effective at identifying providers with similar patterns of claims. The anomaly detection technique is effective at identifying providers with unusual patterns of claims. However, the approach may not be effective for identifying complex fraud patterns that involve multiple providers. The study is limited by the use of a single dataset and the lack of comparison with other unsupervised machine learning methods.

1. **CRISP-DM Approach**

A well-known and frequently employed methodology for data mining and analytics projects is called CRISP-DM (Cross Industry Standard Process for Data Mining). It offers an organized and iterative method to direct every step of the data mining process, from identifying the business issue to implementing the fix. Six key phases make up the CRISP-DM framework, and all that was covered in each of the six phases is outlined below. The CRISP-DM Model that was used is depicted in Figure 1.

***Business Understanding***

The team met in order to discuss various project concepts and objectives throughout the business understanding phase. To build a clearly defined project with specific goals, each team member was encouraged to submit their ideas and undertake literature reviews. The team members expressed enthusiasm for using a variety of machine-learning techniques in the project and suggested their personal interests in potential project topics. Each team member was tasked with doing literature surveys to learn more about potential techniques, methodologies, and technologies that may be used to arrive at an ideal solution when the project topic was decided. During this phase, the team also developed a project plan. The project plan specified the important deadlines, deliverables, and milestones for various phases of the project. It provided a breakdown of the duties and tasks that were given to each team member, facilitating effective coordination and communication.

***Data Understanding***

In the data understanding phase, the team thoroughly explored healthcare claims data obtained from Kaggle. EDA was performed to investigate the variables, size, and structure of the data in order to get insightful knowledge. With regard to missing values, outliers, and inconsistencies, the quality of the data was evaluated. It ensured the relevance of the data to the project's objectives, verifying the presence of necessary variables for detecting healthcare fraud. This thorough comprehension of the data, together with its reliability and applicability, created a solid groundwork for succeeding steps.

***Data Preparation***

The phase of data preparation included a number of crucial procedures to convert the data into a format appropriate for analysis and model training. This stage involved cleaning the data to deal with missing numbers, outliers, and irregularities. Techniques for transforming data are used, such as scaling numerical variables and encoding categorical variables. Feature engineering was performed to produce additional variables that improve predicted accuracy. Additionally, methods like oversampling were used to address data that is unbalanced. The prepared dataset is subsequently divided into training, validation, and testing sets using sklearn’s train\_test\_split. For future data analysis or modeling tasks to be accurate and reliable, the data preparation phase is essential.

***Data Modelling***

The team concentrated on developing and improving machine learning models designed to identify healthcare fraud throughout the data modeling phase. The creation of specific machine learning algorithms like Random Forest, Logistic Regression, Decision Tree, and XGBoost was the first major step in this period. Using the prepared dataset, the models were then trained and validated. To improve model performance and generalization, the team used iterative optimization strategies, fine-tuning hyperparameters using grid search and cross-validation techniques, choosing pertinent features, and using appropriate sampling techniques.

***Data Evaluation***

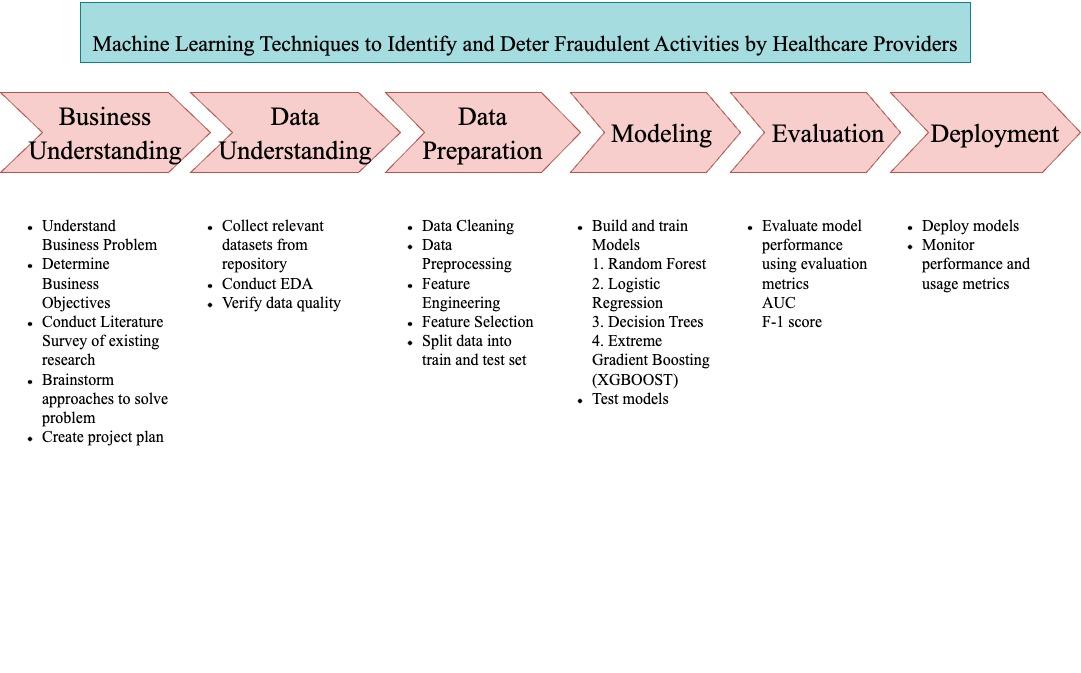
The performance and efficacy of the created machine learning models are the main objectives of this project's data evaluation phase. In order to compare these algorithms' performances and assess how well they can spot fraudulent activity, measures like AUC and F1 scores are taken into account. To improve the performance of the models, fine-tuning techniques are used. These include changing the hyperparameters, feature selection, and sampling procedures. As a result of the evaluation's findings and conclusions, decision-making processes relating to healthcare fraud detection and prevention are guided, allowing for the identification of high-risk regions and the application of successful fraud-reduction and patient safety-improving techniques while also lowering healthcare costs. The accuracy of Decision Tree and XG Boost, 75.25 and 75.19%, respectively, was good.

***Deployment***

Multiple steps make up the deployment process, which is used to operationalize a machine-learning model for usage in a real-world setting. The future scope of this project includes this procedure. It involves the following steps: The trained model is first saved in the proper format, after which infrastructure is created to host the deployed model, guaranteeing scalability and top performance. The goal of data input and output facilitation is integration with current production systems. To prepare and provide data to the model, a data pipeline can be set up. To ensure accuracy and functionality, stringent testing and quality assurance procedures are used. To track performance and deal with any possible problems, ongoing monitoring and maintenance methods are put in place.

**Figure 1**

*CRISP-DM Methodology Followed in the Project*

**

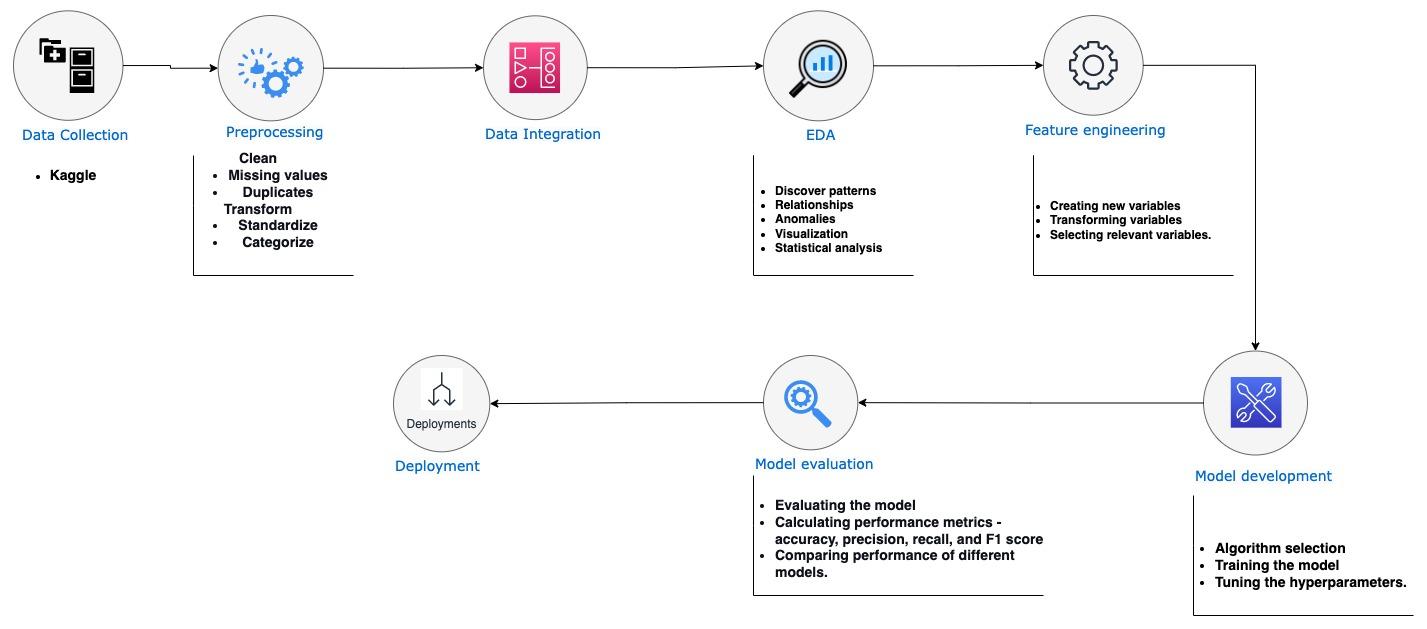
1. **System Architecture**

The healthcare fraud detection project uses a comprehensive system architecture with several stages as shown in the system architecture design in Figure 2. Beginning with the acquisition of legitimate healthcare claims data from Kaggle, data gathering integrity is ensured. The obtained data is next subjected to preprocessing, which includes crucial actions including data cleansing and transformation methods like standardization, normalization, and categorization. These preparation techniques improve the quality of the data and make it easier to extract valuable features. Exploratory Data Analysis (EDA) is used to uncover patterns that are pertinent to fraud detection and obtain an in-depth understanding of the properties of the data. Additionally, relevant variables are created, transformed, and chosen using feature engineering techniques.

The next stage is devoted to developing the models, and it makes use of powerful algorithms like Random Forest, Logistic Regression, Decision Tree, and XGBoost. These algorithms can effectively handle categorical and continuous data, which makes it possible to build reliable fraud detection models. To gauge the models' success in spotting fraudulent activity, performance metrics like accuracy, F1 score, and AUC are used in their training and evaluation. In the future scope of the project, after selecting the optimal models, the focus shifts to deploying them in a production environment, which includes setting up a dedicated infrastructure for hosting the models. Routine updates and changes can take care of any problems and improve overall functioning, ongoing monitoring and maintenance procedures assure the system's dependability and performance. The ultimate objective of the project is to improve patient safety and the efficiency of healthcare initiatives by identifying and preventing fraudulent activities, eliminating offenders, and cutting healthcare expenses.

**Figure 2**

*Healthcare Provider Fraud Detection System Architecture*

****

1. **Data Preparation**

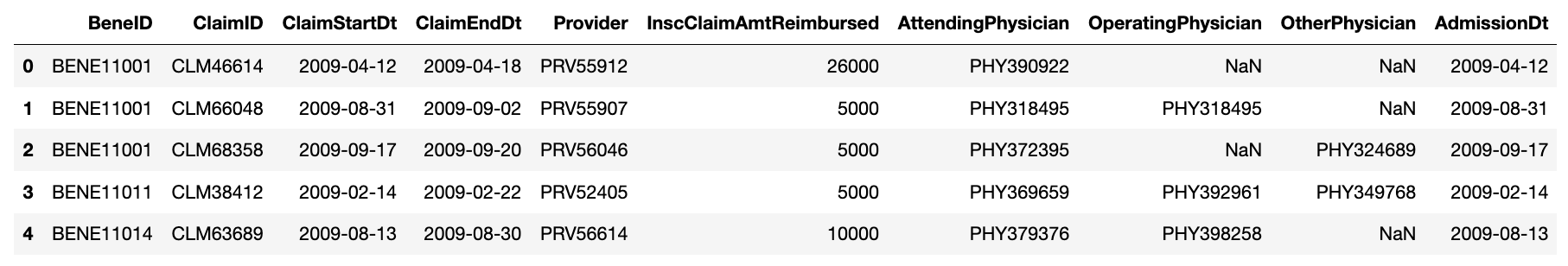
***4.1. Data Exploration***

Understanding the data is the initial stage in the data exploration process. This involves being aware of the many types, qualities, and distributions of data. Inpatient claims, Outpatient claims, and Beneficiary details are the three main formats in which the data for this project is available. Information regarding patients who are admitted to hospitals is available in the inpatient claims data, including the patient's admission and discharge dates, diagnostic codes, and procedures carried out. The information on patients who visit the hospital but are not hospitalized, including their date of service, diagnostic code, and procedures carried out, is provided by the outpatient claims data. Information on the patients, including their age, gender, and medical problems, is included in the beneficiary details data. The figures 3, 4 and 5 below show the inpatient claims data, outpatient claims data, and beneficiary for each provider data.

The dataset files are added to Google Drive and can be found using the below link - <https://drive.google.com/drive/folders/1o6zto5bHZd3eURIdzOkShPAfJLTDfL1a?usp=share_link>

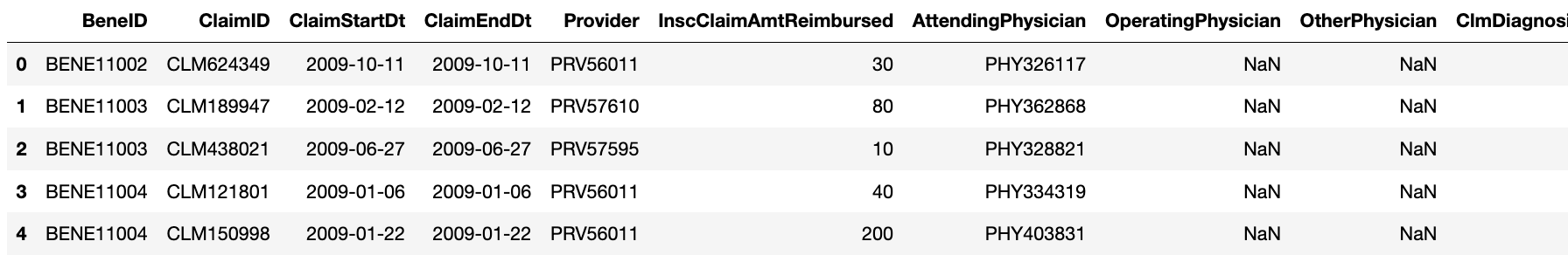
**Figure 3**

*Sample Inpatient Claims Data*



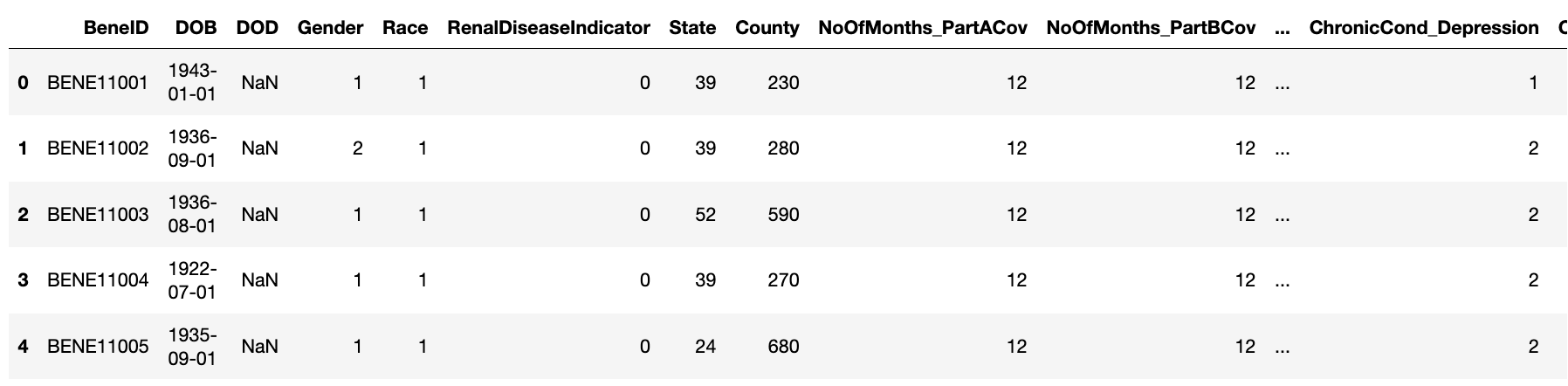
**Figure 4**

*Sample Outpatient Claims Data*

**

**Figure 5**

*Sample Beneficiary Data*

**

Creating an ideal model to forecast fraudulent activity is the next step once possible patterns have been found. For this objective, a number of machine learning methods can be applied. The ideal algorithm will vary depending on the particular characteristics of the data. The patient's health problems, the provider's claim history, and the price of the services offered are likely to be the factors in this scenario that are most crucial.

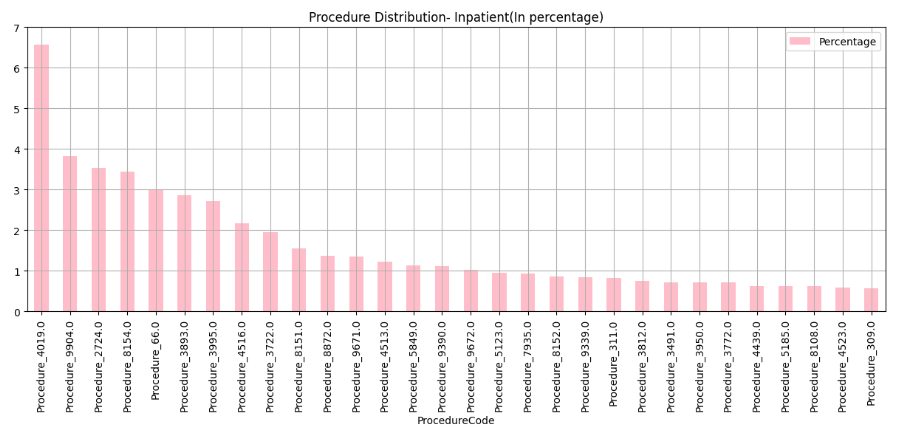
The development of an accurate model to predict fraudulent behavior is an important step in protecting Medicare and other healthcare programs from financial loss. Once an ideal model is created, it can be used to identify which providers are more likely to be fraudulent. These providers can then be investigated further to determine if they are actually committing fraud.

One way to identify potential fraudulent providers is to look at the procedures they perform. As shown in Figure 6, procedure code 4019 is the most common procedure performed on inpatients. This procedure is used to treat a variety of conditions, including hernias, ulcers, and tumors. It is important to note that not all providers who perform procedure code 4019 are fraudulent. However, it is a red flag that should be investigated further. General surgery procedure code 4019 is used to treat a number of ailments, such as hernias, ulcers, and tumors. A diagnostic test called 9904 is used to assess how well the heart and lungs are working. A kidney stone can be surgically removed using procedure code 2724. Diabetes, heart disease, and cancer are just a few of the ailments that are treated by medical procedure code 8154. A diagnostic test called procedure code 66 is used to assess how well the nervous system and the brain are working.

The high incidence of procedure code 4019 indicates that it is a widely used and successful treatment for a number of illnesses. The fact that the other top four procedure codes are utilized to treat a wide range of ailments implies that inpatients are frequently sent to the hospital for complicated and critical medical issues.

**Figure 6**

*Procedures Distributions in Inpatient Data*

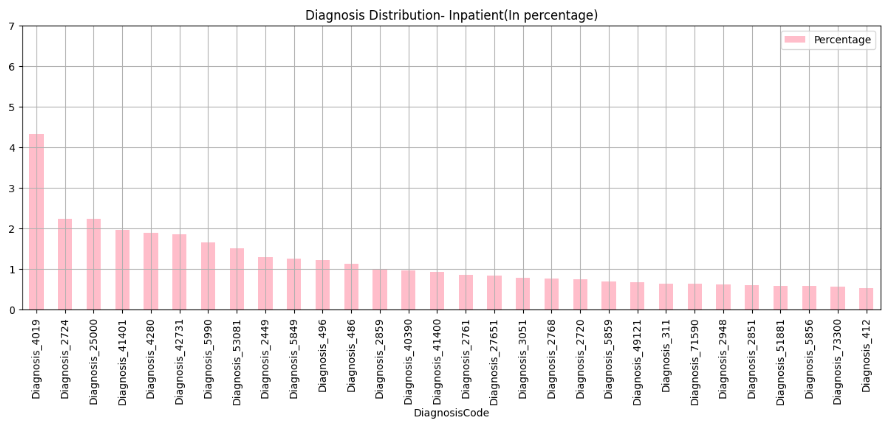
**

Furthermore, unspecified essential hypertension, or diagnostic code 4019, is the one that patients receive the most frequently as seen inFigure 7. In 4.5% of cases, this diagnosis has been made. For inpatient data, the top 5 diagnostic codes are 4019, 2724, 25000, 41401, and 4280.

A chronic disease that affects the heart and blood vessels is identified by the diagnosis code 4019. High blood pressure, which can cause heart disease, strokes, and other health issues, is one of its defining characteristics. Pneumonia is a lung ailment, and its diagnosis code is 2724. A fracture, often known as a break in a bone, is diagnosed with the code 25000.

A heart attack, also known as an abrupt decrease of cardiac blood flow, is designated by the diagnosis number 41401. A stroke is characterized by an abrupt reduction in blood flow to the brain and has the diagnosis code 4280. Since these disorders are severe and necessitate hospitalization, they are the most prevalent diagnosis for inpatient data. The cost of treating them might also be high. It's critical to be aware of these disorders so that you can take precautions against them or seek treatment if they manifest.

**Figure 7**

*Diagnosis Distribution in Inpatient Data*

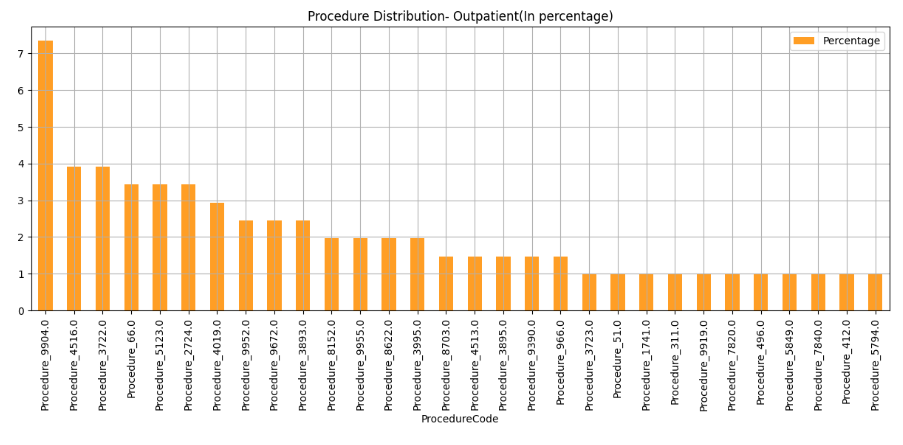
The procedure code 9904, which represents 7.5% of all operations, is the most popular for outpatients. For inpatients, the top 5 procedure codes are 9904, 3722, 4516, 2724, and 66. According to these findings, general medical treatments are the most often done operations on outpatients, whereas surgical procedures are the most frequently performed procedures on inpatients. This is probably because inpatients frequently have more severe illnesses than outpatients and need more involved medical attention.

Given that 9904 is the most often used procedure code for both inpatients and outpatients, it is likely a fairly adaptable treatment that may be utilized to treat a wide range of illnesses. The fact that 9904 is a generic medical process that may be used to identify and treat a wide range of medical issues is probably to blame for this.

Overall, the findings from Figure 8 offers useful knowledge on the most frequent procedures carried out on both inpatients and outpatients. Utilizing this information will help patients receive higher-quality treatment and point out places where resources may be spent more effectively.

**Figure 8**

*Procedures Distributions in Outpatient Data*

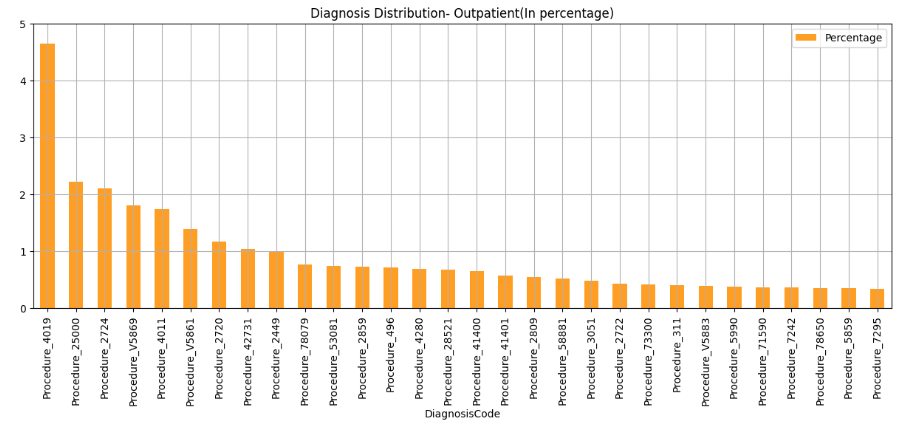
**

Moving on to Outpatient data, the most frequent diagnostic code is 4019, which stands for nonspecific essential hypertension as seen in Figure 9. 4.8% of patients were given this diagnosis. For inpatients, the top five diagnostic codes are 4019, 25000, 2724, V5869, and 401. These codes stand for a number of ailments, such as heart failure, pneumonia, hypertension, and other illnesses.

These results imply that hypertension is a significant health issue for both outpatients and hospital patients. In order to treat and manage patients appropriately, healthcare professionals must be informed of the prevalence of this disorder.

**Figure 9**

*Diagnosis Distribution in Outpatient Data*

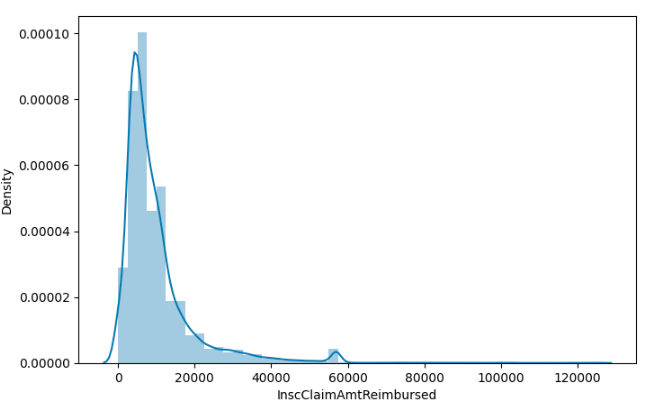
**

In Figure 10, the amount provided as claim reimbursement appears to be distributed in a lognormal manner. This means that the majority of claims receive little amounts of money in reimbursement, whereas a small number of claims receive significant sums of money in reimbursement. The most often amount refunded is zero since the mode of the distribution is at 0. The distribution's mean is around 5000, while its standard deviation is roughly 10000. This indicates that most claims receive payments in the range of $0 to $10,000, with a tiny minority receiving payments in the range of $10,000 to $20,000.

The majority of the reimbursements fall between the range of 0 and 20000. Rarely are sums more than $20,000 received as claim reimbursement. This implies that the bulk of claims relate to relatively common medical treatments, whereas a smaller percentage of claims relate to more involved or expensive procedures. The type of medical operation, the seriousness of the patient's illness, and the hospital's location are just a few examples of variables that may have an impact on how the reimbursed amounts are distributed.

**Figure 10**

*Distribution of Claim Amount Reimbursement in Inpatient Data*

**

The value of 3500 in the 99.9 percentile implies that 3500 is less than or equal to 99.9% of the claim amounts reimbursed for outpatient services. This indicates that less than 0.1% of claim amounts exceed $3,500. The graphic shows that this column has several outliers, or data points that are beyond the 99.9 percentile range. These anomalies might be the result of a variety of things, including pricey medical treatments or equipment or false claims.

A helpful measure for spotting anomalies in the ClaimAmt Reimbursed column is the 99.9 percentile value of 3500. Numerous things might lead to outliers, therefore it's critical to look into them to see if they're real or false. For instance, a costly medical operation, like a heart transplant, can be the reason why an outlier occurred. An outlier, however, could also be the consequence of dishonest behavior, such as making a bogus claim for compensation. To make sure that the data is correct and that fraud is not taking place, it is crucial to look at outliers.

**Figure 11**

*Distribution of Claim Amount Reimbursement in Outpatient Data*

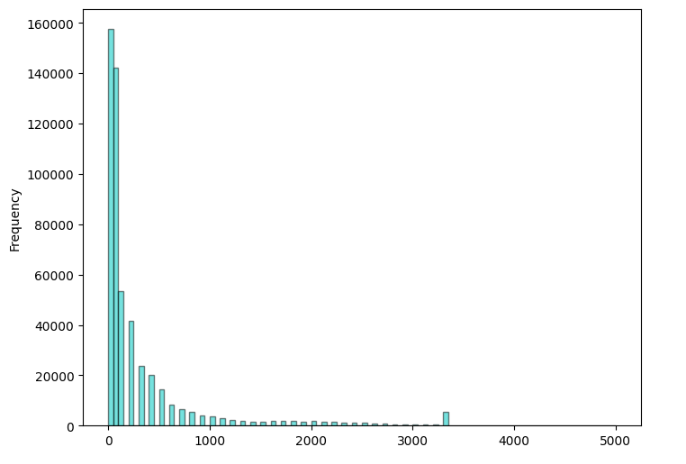
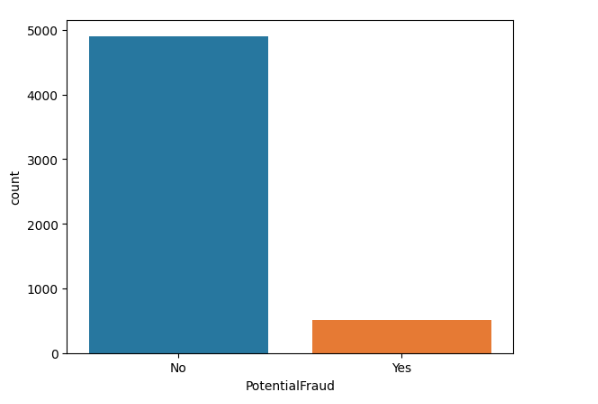


Figure 12 displays an unbalanced dataset. This indicates that one class has more data points than the other class. In this instance, the "No" class has more data points than the "Yes" class. Machine learning models may have issues with this since they may be biased toward the dominant class.

**Figure 12**

*Distribution of Target Variable in Outpatient Data*



The proportion of fraudulent contacts in both inpatient and outpatient settings is shown in Figure 13. The dataset is (23402, 31) in form, and 57.81% of the fraud instances include inpatient records. Accordingly, about 58% of the inpatient data used for training contains bogus entries. According to the findings, inpatient settings are more likely than outpatient ones to be linked to fraudulent activities. There are several reasons why this can be the case. First, compared to outpatient settings, inpatient settings are sometimes more complicated and expensive. They become more desirable targets as a result for scammers. Second, patient information is frequently more accessible in inpatient settings, which makes it easier to conduct fraud. Last but not least, inpatient settings frequently have more regulations than outpatient ones, which might make it more challenging to identify and stop fraud.

**Figure 13**

*Checking for Fraudulent Claims in Inpatient Data*

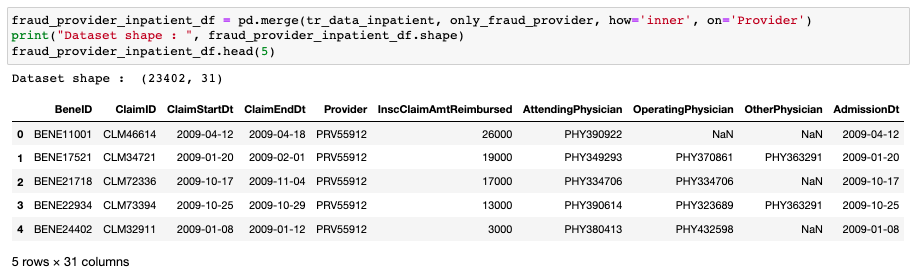
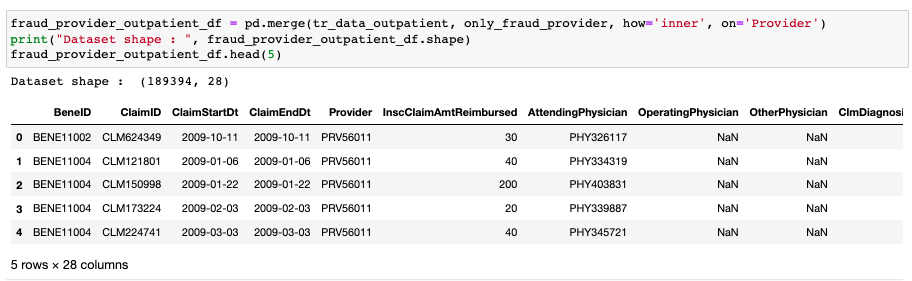
**

Figure 14 displays the dataset structure and the percentage of outpatient data-related fraud instances. There are 28 columns and 189,394 rows in the dataset. 36.58% of these situations involve information from outpatient patients. This indicates that around 69,312 of the dataset's instances are associated with outpatient data. 37% of these are fake entries. This implies that about 25,545 of the dataset's outpatient cases are fake.

Concern should be expressed over the significant number of false entries found in the outpatient data. This implies that widespread fraud is taking place in the healthcare industry. The victims of this scam may include both patients and service providers. Patients may have their valid claims refused, and providers may be required to pay for false claims. It is crucial to take action to combat this fraud, including putting fraud detection and prevention mechanisms in place.

**Figure 14***Checking for Fraudulent Claims in Outpatient Data*

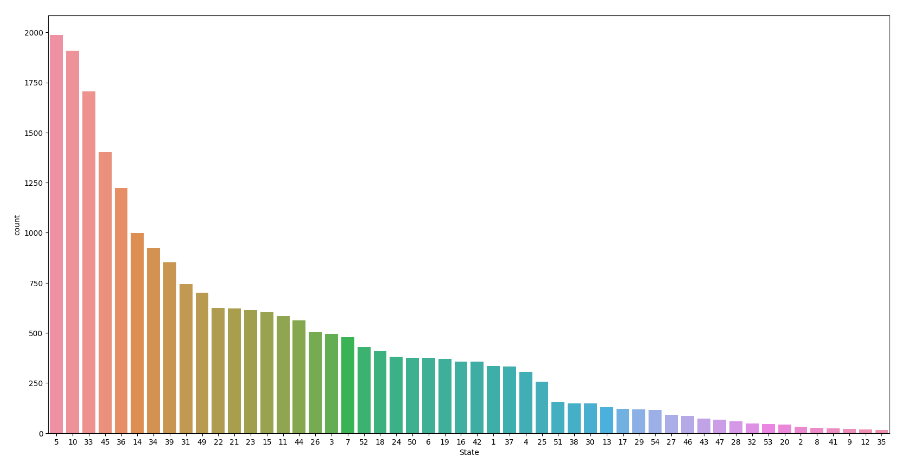


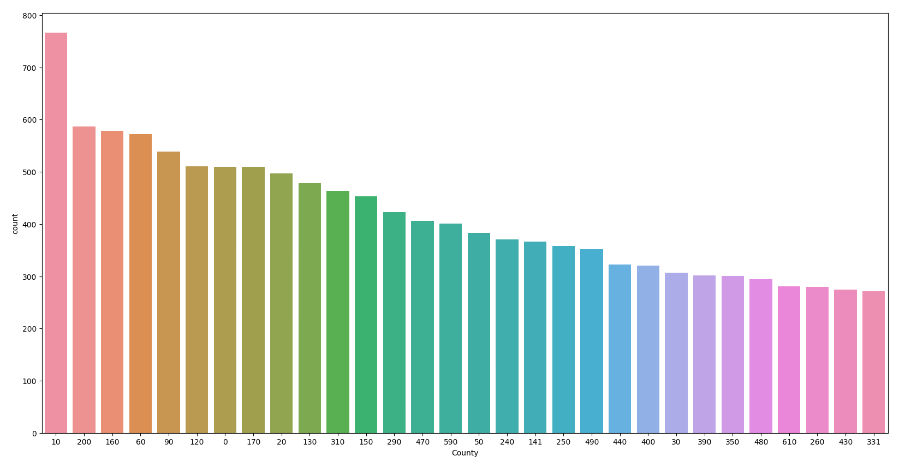
According to Figure 15, the states and counties with the highest number of fraudulent inpatient visits had the following codes: 5, 10, 33, and 45. A combined total of 62% of all fraudulent inpatient encounters occur in these states and counties. The prevalence of large urban areas with high rates of uninsured and underinsured people, a large number of hospitals and other healthcare facilities, a large population of Medicare and Medicaid recipients, and a large number of healthcare fraud schemes may all contribute to the high rates of fraudulent inpatient encounters in these regions.

These regions have a severe issue with high rates of fraudulent inpatient visits that costs the taxpayers billions of dollars annually. By giving patients unneeded or subpar care, these fraudulent interactions also put patients at risk. The issue of fraudulent inpatient encounters can be solved in a number of ways, including increasing the quantity of inpatient claim audits, stiffening penalties for fraudulent inpatient claims, encouraging more people to come forward and report fraudulent inpatient claims, and educating healthcare professionals more about the dangers of such claims. By doing these things, we may lessen the number of fraudulent inpatient encounters and safeguard patients.

**Figure 15**

*States and Counties with the Highest Number of Fraudulent Inpatient Visits*

**

**

Further analysis indicates that around $290 million was lost to fraud in total in 2009. Several tactics were used to lose this money, including invoicing for services that were not provided, making fraudulent claims, and overcharging for services. $241 million was lost in the inpatient environment, where money was lost most frequently. The outpatient setting was where the remaining $54 million was squandered.

A significant issue is the widespread fraud in the healthcare sector. Instead of being wasted to scammers, this money may be utilized to treat patients. Fraud is being fought against by the government and healthcare professionals, but it is a challenging issue to resolve. Healthcare providers may contribute to the reduction of fraud by putting in place robust internal controls and being watchful in spotting questionable conduct.

***4.2. Data Pre-Processing***

Each of the data files' columns has undergone exploratory data analysis (EDA). Variable relationships as well as data kinds, distributions, and types of data have all been studied. Additionally, missing data and outliers have been found.

We'll now combine all of the data files into one dataset. This will make it possible to do more complex EDA, such as correlation analysis and cluster analysis. This combined dataset may be used to create machine-learning models. After combining and doing the target variable analysis; there were 345,415 "No" cases, which is a lot more than there were "Yes" cases (212,796).

There are no discernible differences in the age distribution of the patients who submitted claims that may be utilized to detect fraud. The number of prospective fraud cases involving people over 65 is, however, on the rise. Additionally, this age group is represented by the majority of the patients who submitted claims.

This implies that there may be a link between age and the likelihood of fraud. It's probable that due to their advanced age and pre-existing medical issues, older people are more prone to become victims of fraud or to be the targets of fraud. The same has been shown in Figure 16.

**Figure 16**

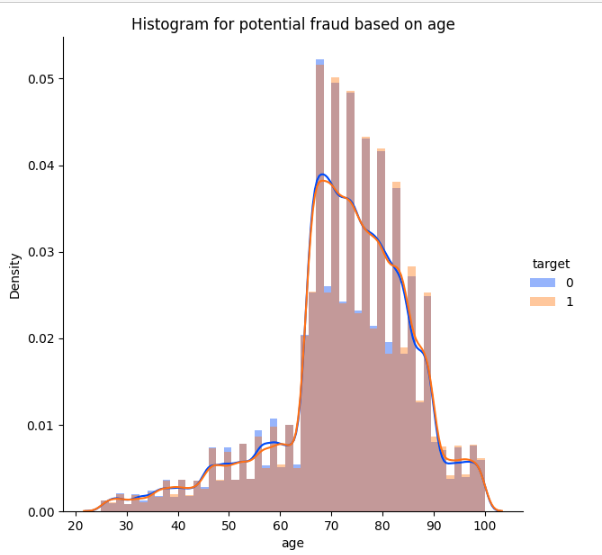
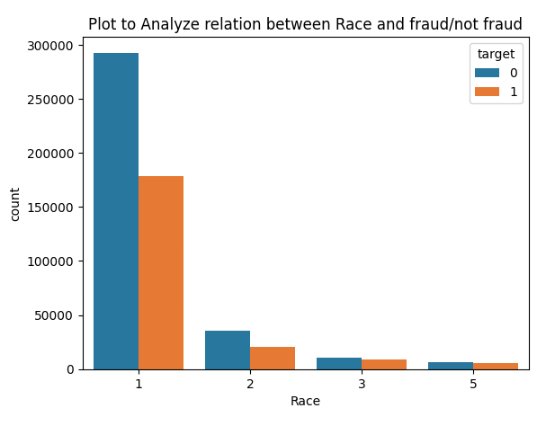
*Age Distribution Among the Claims  
*

Figure 17 below demonstrates that patients of a certain race, shown as 1, account for the bulk of fraudulent instances. This is a worrying discovery because it raises the possibility that the healthcare system has a bias against this particular race. To comprehend the underlying reasons behind this bias and to take action to alleviate them, it is crucial to look into this issue further.

There are several reasons why patients of race 1 may experience more fraudulent cases than other patients. One explanation is that scammers are more likely to target them. Another hypothesis is that they could have less access to high-quality medical treatment, which might leave them more open to fraud.

**Figure 17**

*Race Distribution Among the Total Claims*

**

***4.3. Data Transformation***

The process of transforming or altering data from one format or structure to another with the aim of improving its usability and application for diverse purposes is referred to as data transformation. Data must be changed in order for it to be compatible with other systems, programs, or analytical methods. Data cleansing, filtering, aggregation, normalization, and encoding are just a few of the many techniques that fall under the category of data transformation. By eliminating duplications and standardizing variables, these transformations can enhance the quality of the data and make it easier to analyze, visualize, and make decisions. Organizations may fully utilize their data and get insightful information that can spur innovation, efficiency, and well-informed decision-making by implementing the right data transformation techniques.

By deducting the admission date from the discharge date, it is possible to determine the total number of days a patient was admitted to the hospital. We can continue to use 0 for the number of days admitted for outpatients who don't spend the night in the hospital. A patient would have spent two days in the hospital, for instance, if they were admitted on May 18, 2023, then released on May 20, 2023. The number of days hospitalized would be zero if a patient was admitted on May 18, 2023, and discharged the same day.

***4.4. Data Normalization***

The problem of variable sizes and distributions within datasets is addressed by data normalization, a fundamental preprocessing step in data analysis and modeling. It is essential for delivering dependable and accurate outcomes. The value of data normalization rests in its capacity to make the data consistent and comparable, allowing for fair judgments and objective analysis.

Data normalization has a variety of effects on the model's performance. First off, it aids in removing the dominance of some characteristics or variables as a result of their greater scales or wider ranges. Normalization prevents any one feature from overpowering others by scaling all features to a common range, which enables the model to weigh each feature's contribution more equally. This helps prevent biased outcomes and guarantees a fair representation of all factors.

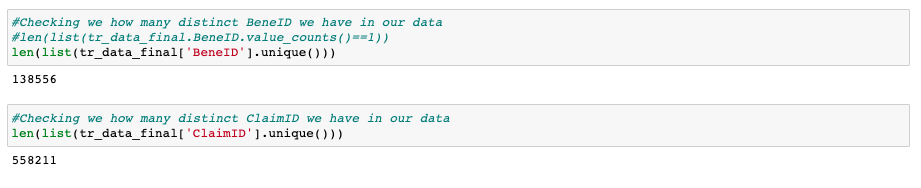
Normalization also aids in enhancing convergence during model training. Numerous optimization methods and machine learning algorithms rely on distance computations or gradient descent algorithms, which are sensitive to the input data's size. Significantly differing sizes between features might cause the model to converge slowly or possibly not at all. By bringing all characteristics to a comparable scale through data normalization, this problem is solved, allowing for effective optimization and quicker model convergence.

The collection contains 138,556 distinct BeneIDs, as shown in Figure 15. This indicates that the dataset has 138,556 unique recipients. The fact that there are 558,211 unique ClaimsIDs in the dataset suggests that each recipient may have submitted claims more than once. This implies that a distinct record exists for each time a beneficiary submitted a claim. The absence of duplicate rows in the dataset shows that the information is accurate and unaltered. This is significant since it shows that the information may be relied upon and applied to the analysis.

The information in Figure 18 may be utilized to better understand the claims procedure and pinpoint areas that could require improvement. For instance, the existence of 138,556 distinct BeneIDs indicates that there are many beneficiaries in the system. This indicates that it's critical to provide beneficiaries with a simple and effective claims process. There are 558,211 unique ClaimsIDs, which shows that recipients may submit claims more than once. As a result, it's critical to keep track of claims and make sure beneficiaries aren't being overcharged.

**Figure 18**

*Insights into the Claims Process*

**

The number of distinct doctors for each patient is shown in Figure 19**.** This data may be utilized to comprehend the patient's medical team and spot any treatment gaps. Connecting a patient with additional experts may be crucial, for instance, if they have only seen one doctor thus far. For simpler analysis, the different sorts of doctors can be expressed as numerical numbers. To do this, a lookup table that associates a different number with each variety of doctor may be made. Once the different sorts of doctors have been encoded, they may be utilized to respond to inquiries regarding the patient's medical staff. One can find out, for instance, how many different kinds of doctors have treated a patient.

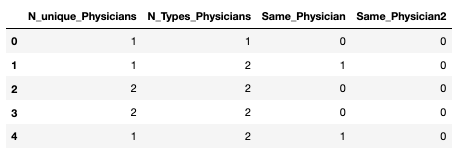
A patient's healthcare experience may be understood by counting the number of different doctors who treat them. A clue that a patient is not getting coordinated treatment, for instance, is if they have seen several different doctors. Look for patients who have just one distinct physician listed in their records to find individuals who were only seen by that doctor. Because they might not have access to a variety of experts, these patients run the risk of not getting the best care available.

The frequency with which each physician appears in the patient's record can be used to establish whether one doctor is serving in more than one capacity to care for a patient. If a doctor is included in the patient's record more than once, it is likely that they are involved in the patient's treatment in more than one capacity. A doctor could do both the patient's surgery and primary care, for instance.

The knowledge from Figure 19 may be applied to raise the standard of patient care. It is feasible to spot treatment gaps and link patients with the appropriate specialists by having a thorough grasp of the patient's healthcare team. A patient's healthcare experience may also be evaluated by knowing the different sorts of doctors that have treated them. Finally, it is feasible to connect patients with additional healthcare professionals by identifying those who have only visited one doctor.

**Figure 19**

*Calculated Fields to Study the Impact of Physicians*

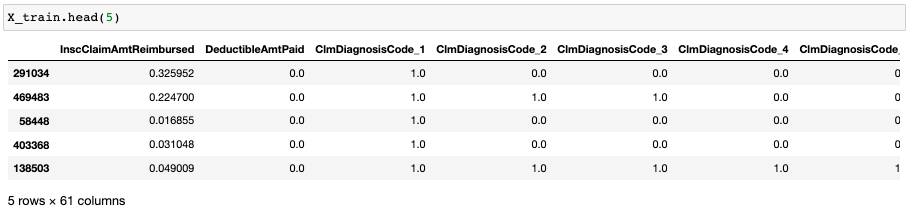
**

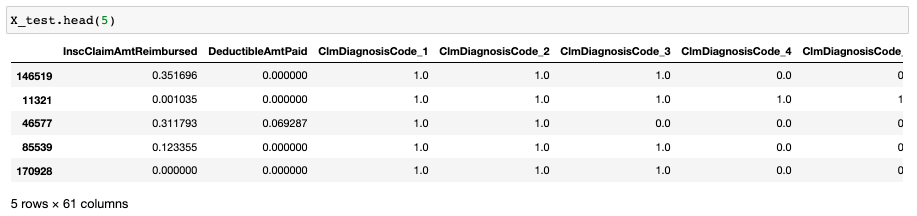
The training and test sets' columns are normalized using the aforementioned algorithm. This is done to make sure that every column has the same scale, which can help machine learning models perform better. The Normalizer class from the sklearn.preprocessing module is initially imported by the code. Then, num\_col\_normalizer(), a function with three inputs (X\_tr, X\_te, and cols), is defined. The training and test sets are represented by the variables X\_tr and X\_te, respectively, while the list of column names to be normalized is represented by the argument *cols*.

A Normalizer object is initially created using the num\_col\_normalizer() method. The Normalizer object is then fitted to the training set using the cols parameter. It then applies the Normalizer object to convert the training and test sets. The normalized training and test sets are then returned. The list of columns to normalize is then defined as cols\_to\_normalize.

**Figure 20**

*Normalized Train and Test Dataset Preview Respectively*

**

**

Further, as part of the data normalization and model development, we define a few functions:

* For a given collection of true labels y and predicted probabilities y\_pred\_prob, the function get\_threshould\_fpr\_tpr(y, y\_pred\_prob) computes the false positive rate (FPR), true positive rate (TPR), and thresholds using the roc\_curve function.
* The results for the train and test FPR and TPR are used to create a ROC-AUC plot using the function draw\_Roc\_Auc\_plot(train\_fpr, train\_tpr, test\_fpr, test\_tpr). It displays a visual representation of the model's performance by plotting the AUC (Area Under the Curve) for both the train data and the test data.
* By locating the highest value of tpr\*(1-fpr) and the related threshold, the function find\_best\_threshold(threshold, fpr, tpr) determines the best threshold. The projected probabilities must meet this level in order to be divided into binary groups.
* Based on the projected probabilities and the best threshold, the function predict\_with\_best\_t(probe, threshold) forecasts the binary class labels. If the anticipated probability exceeds or equals the threshold, a value of 1 is assigned; otherwise, a value of 0 is given.
* The confusion matrix is produced by the function get\_confusion\_matrix(y, y\_pred, axis, best\_t) using the true label y and the predicted label y\_pred. Using Seaborn, it shows the confusion matrix as a heatmap.
* The model's performance is assessed using the function model\_performence\_check(model, X\_train, X\_test, y\_train, y\_test). It calculates the predicted probabilities for the train and tests data using the provided model. Then, for both sets of data, it obtains the FPR TPR, and thresholds. It displays the confusion matrices for the train and test data, plots the ROC-AUC curve, and chooses the best threshold. The F1 score and AUC score of the model for the test data are then computed and returned.

1. **Model Selection**

The project creates models that identify medical service providers who might be submitting false claims. Several different machine learning techniques, such as Logistic Regression, Random Forest, Decision trees, and XGBoost, were examined for model selection in order to reach this goal. These models are evaluated using several different performance standards, including the confusion matrix, precision, recall, area under the curve (AUC), and F1 score. Using these indicators, we can determine how well the algorithms can distinguish between legitimate and counterfeit healthcare providers. The class imbalance was taken into consideration in the model's selection process because there are more providers in the dataset than fraudulent ones. Improper fraud detection can result from applying a model that is biased against the dominant group because of an imbalanced dataset.

**5.1 Model Implemented**

***Logistic Regression***

One common application of logistic regression is in the prediction of one of two classes of outcomes in binary classification issues. Healthcare experts' claims may reveal their accuracy. Logistic Regression uses a logistic function to show the relationship between independent variables (features) and outcome likelihood. This function converts the linear combination of characteristics into a probability interval between 0 and 1. The model shows each feature's fraud prediction value by calculating its coefficients. The coefficient represents the logarithmic change in the outcome if the relevant attribute is increased by one unit. The coefficient has a positive or negative link with fraud. The model was implemented in the project and the results of the model can be seen in Figure 21.

**Figure 21**

*Plot and Results for Logistic Regression Model*

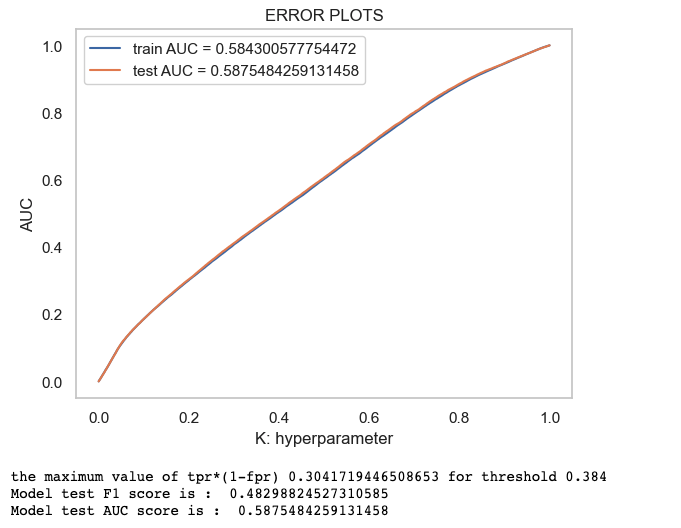
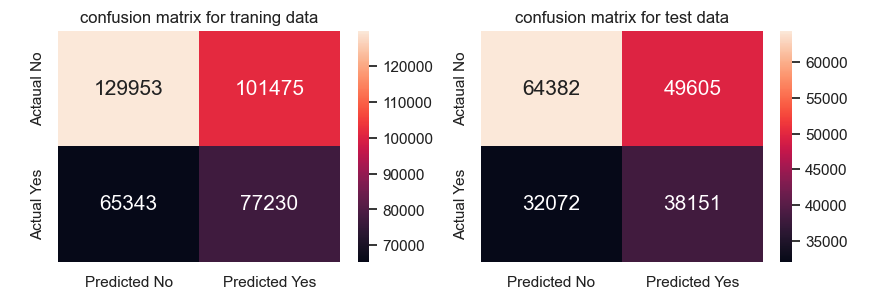


Figure 22 shows the confusion matrix on training data and testing data for logistic regression.

**Figure 22**

*Confusion Matrix for Training and Test Data (Logistic Regression Model)*



***Decision Trees***

Decision Trees are machine learning models that use a tree-like structure to produce predictions by recursively dividing the data based on different parameters. Each internal node of the tree represents a test on a particular feature, and each leaf node represents a predicted outcome or class. The model's decision-making process is easier to understand because of the tree structure. From the root node to the leaf nodes, the outcomes of feature testing can be viewed. Decision Trees' open structure makes them useful for spotting indicators of fraud committed by healthcare providers. Decision Trees can capture nonlinear features-target variable relationships. They can divide the data into fraud-related subsets by splitting it by features. Decision Trees are flexible enough to handle complicated patterns and feature interactions. This model is more prone to overfitting when trees are too deep and complex. When the model overfits, it fails to generalize to fresh data. The outcomes of using this model can be seen in Figure 23.

**Figure 23**

*Plot and Result for Decision Trees*

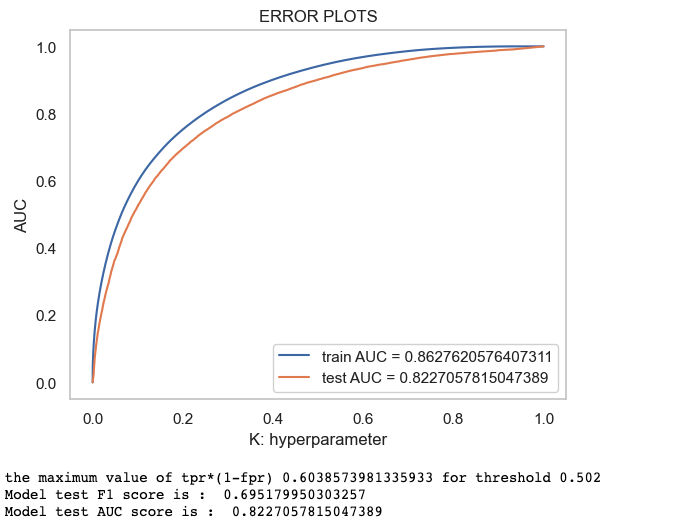
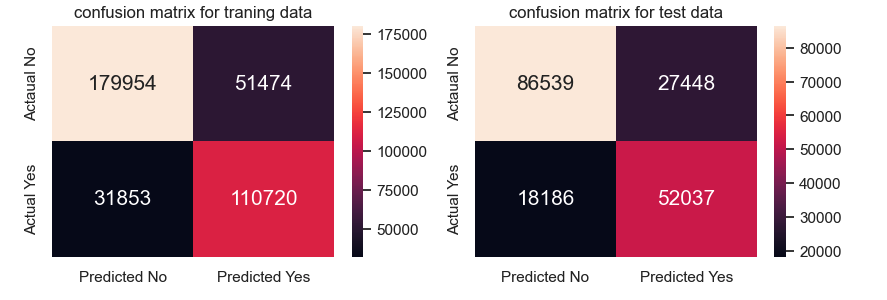


Figure 24. shows the confusion matrix on training data and testing data for decision trees.

**Figure 24**

*Confusion Matrix for Training and Test Data (Decision Tree Model)*



***Random Forest***

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is known for its ability to improve the performance and robustness of individual decision trees. Through the use of a method known as bootstrap sampling, it generates a number of distinct subsets from the initial dataset. Each subset is produced by randomly sampling the data and then replacing some of the samples. This indicates that certain observations may appear more than once in a subset, while others may not be included in the analysis at all. Random Forest accomplishes this variation among the different trees through the process of constructing these subsets. When building each decision tree, Random Forest also randomly selects a subset of features at each node in addition to doing bootstrap sampling as described before. This random feature selection ensures that each tree is trained on a separate set of characteristics, which reduces the likelihood of the trees being highly associated with one another and makes the ensemble more diverse. Figure 25 illustrates the results that were predicted after applying this model.

**Figure 25***Plot and Results for Random Forest*

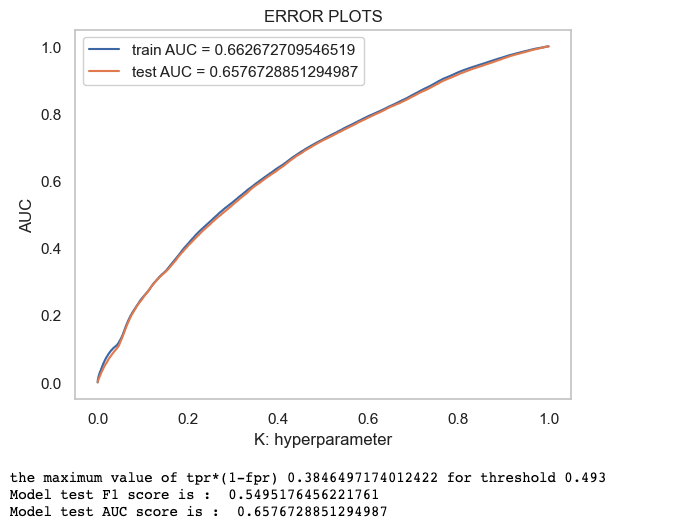
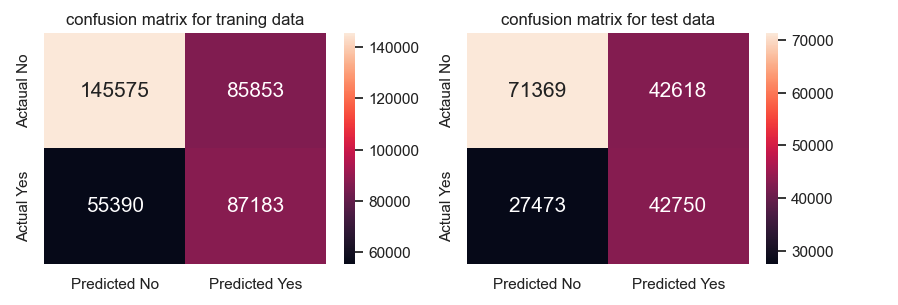
****

Figure 26. shows the confusion matrix on training data and testing data for Random Forest Model.

**Figure 26***Confusion Matrix for Training and Test Data (Random Forest Model)*

****

***XGBoost***

XGBoost is a cutting-edge implementation of the gradient-boosting algorithm family. It is well-known for its outstanding performance in a variety of machine learning applications, including the identification of fraud in healthcare provider claims. XGBoost uses ensemble learning to generate a powerful predictive model from many weak prediction models, usually decision trees. XGBoost improves accuracy and efficiency over gradient boosting methods. In addition, one of the most important features of XGBoost is its capacity to deal with distorted datasets, which are frequently encountered in the field of fraud detection. Adjusting class weights or employing subsampling strategies are two of the options that are available to users of XGBoost in order to overcome the class imbalance.

These methods serve to ensure that the model is not biased against the majority class and accounts for the disproportionately large number of legitimate service providers. In addition, they help to account for the disproportionately small number of fraudulent service providers. The model was implemented in the project and the results of the model can be seen in Figure 27.   
**Figure 27**

*Plot and Results for the XG-Boost Model*

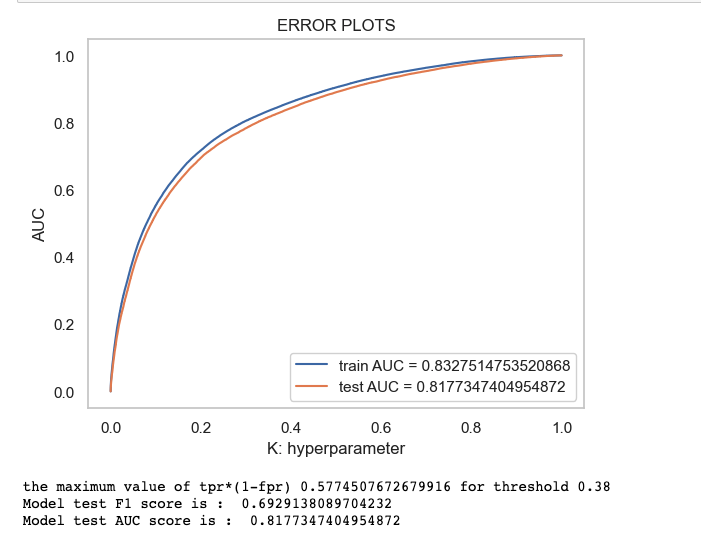
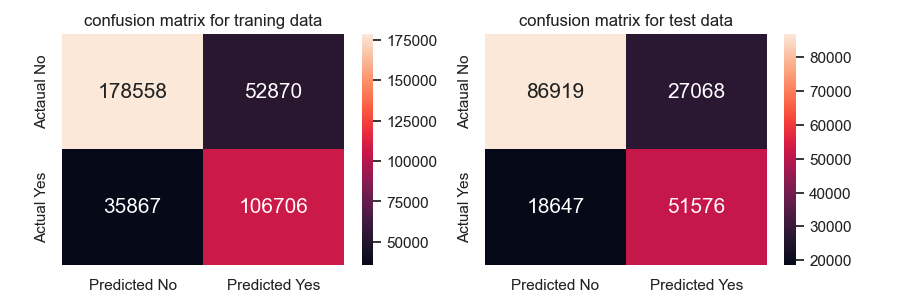


Figure 28 shows the confusion matrix on training data and testing data for XG-boost.

**FIGURE 28***Confusion Matrix for Training and Test Data (XG-Boost Model Model)*



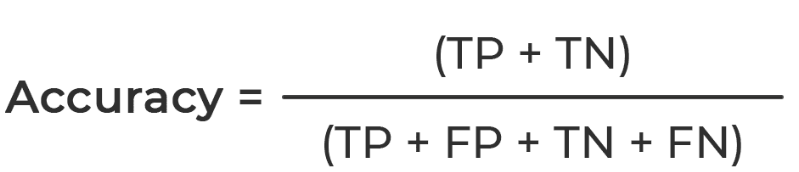
***5.2 Training Model and Results***

***Accuracy***

Accuracy is a commonly used evaluation metric that measures the overall correctness of predictions made by a model. It calculates the ratio of correct predictions to the total number of predictions. It provides a straightforward measure of how well the model performs in terms of correctly classifying instances. It is suitable for balanced datasets where the number of instances in each class is similar. Figure 29 displays the accuracy formula.

**Figure 29**

*Accuracy Formula*

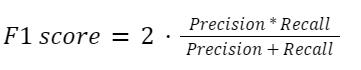


***F1 Score***

The F1 score is a balanced evaluation of a model's performance that includes both precision and recall. It takes into account both the accuracy with which positive instances can be identified (precision) and the completeness with which they can be captured (recall). Precision, or positive predictive value, is a metric used to assess how many positive forecasts actually come to actuality. The recall rate quantifies the percentage of correct predictions made relative to the total number of positive occurrences. An F1 score of 1 indicates perfect precision and recall, while a score of 0 indicates a poor score in either domain. Since it takes into account both the false positive and false negative rates, the F1 score is helpful when working with datasets that are not evenly distributed. Figure 30 shows the F1 score formula.

**Figure 30**

*F1 Score Formula*

**

***Confusion Matrix***

A confusion matrix is a table that summarizes a classification model's performance by showing the numbers of correct, incorrect, and null predictions for each outcome. It enables for the computation of several evaluation measures and provides a full breakdown of the model's performance in each class. In Table 1 all the models are compared on evaluation metrics such as F1, AUC, and accuracy.

**Table 1**

*Comparison Using all Features*

| **Model** | **Hyperparameter** | **Accuracy of test data** | **F1 on Test** | **AUC** |
| --- | --- | --- | --- | --- |
| Logistic Regression | Penalty ‘l2’  C = 10.0 | 0.6298 | 0.4829 | 0.5875 |
| Decision Tree | 'max\_depth': 50, 'min\_samples\_split': 270 | 0.7522 | 0.6951 | 0.8227 |
| Random Forest | 'criterion': 'gini', 'max\_depth': 8, 'max\_features': 'auto', 'n\_estimators': 300} | 0.6387 | 0.5495 | 0.6576 |
| XG Boost | {'n\_estimators': 100, 'eta': 0.3} | 0.7623 | 0.6929 | 0.8177 |

One of the primary objectives of the project was to determine the most significant features for use in predicting healthcare provider fraud. A bar plot was made to show the top attributes in descending order of relevance to provide a visual representation of their significance. The layout made it easy to see how several factors contributed to the whole, so the most important ones could be isolated. This data was crucial in moving the project forward and making informed decisions.

The project team would have been better served if they had known which features were the most crucial before beginning feature selection. This made the model more efficient and understandable by removing superfluous or redundant variables. Furthermore, the data's underlying patterns and correlations could be better understood by gaining an understanding of the primary motivations of fraudulent conduct in healthcare claims. With this information, the team was better able to explain the model's predictions and locate insights that could be put into action to spot and stop fraudulent actions. Figure 31 shows all the important features of the model.

**Figure 31***Feature Importance Chart*

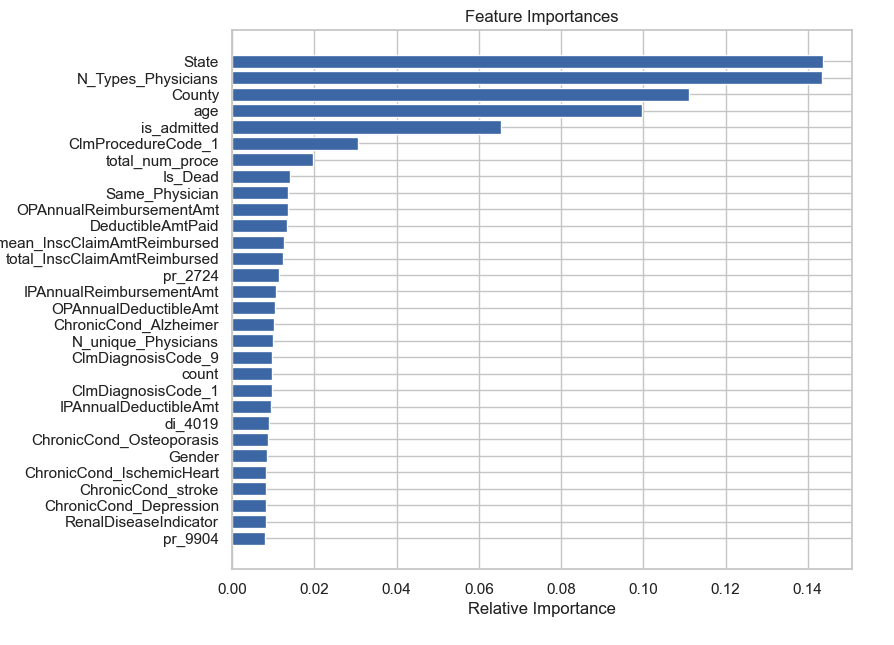
****

Figure 32 shows the model performance summary while using all the features. **Figure 32**

****

The next stage after selecting the project's primary features was to carry out hyperparameter modification with those characteristics. The goal was to improve the model's performance, computational simplicity, and human interpretability by focusing on these essential factors. All models were used on the data which included only important features. Figure 33 shows the output and results of the Logistic Regression model.

**Figure 33**

*Results for Logistic Regression (using important features)*

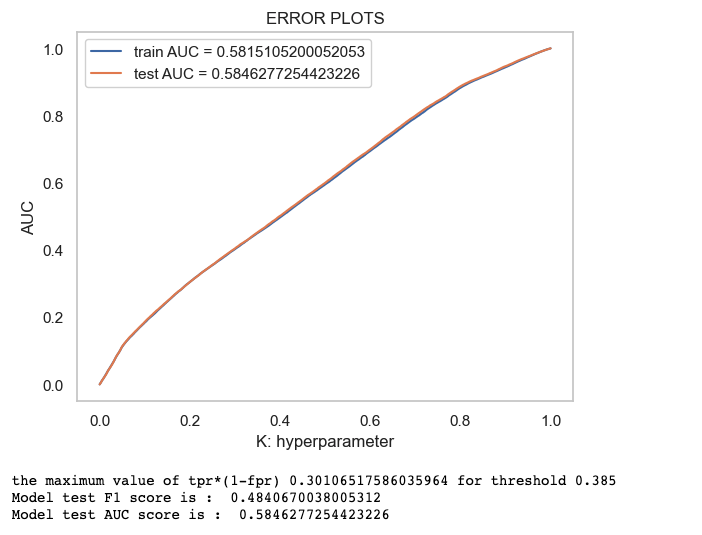
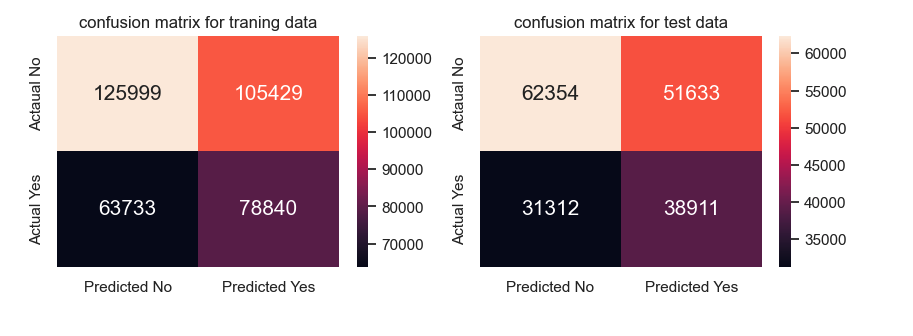


Figure 34 shows the confusion matrix on training data and testing data for Logistic Regression Model just using important features.

**Figure 34***Confusion Matrix for Logistic Regression (using important features)*



**Figure 35** shows the results after using the decision tree model with just important features.

**Figure 35***Results for Decision Trees (using important features)*

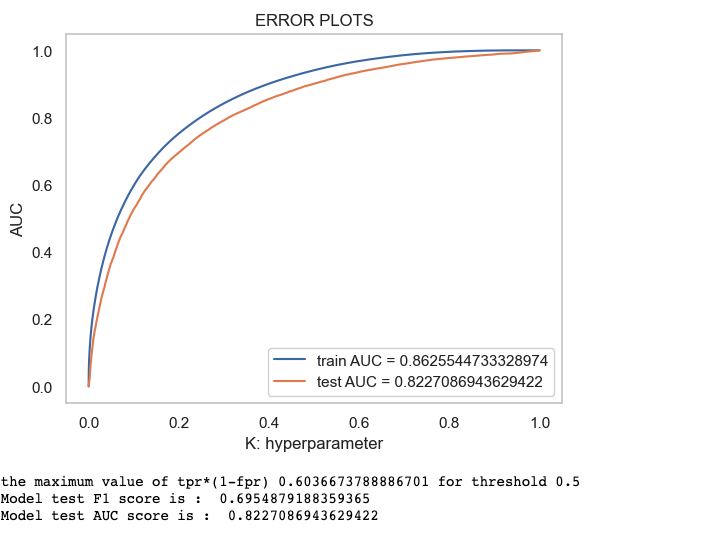
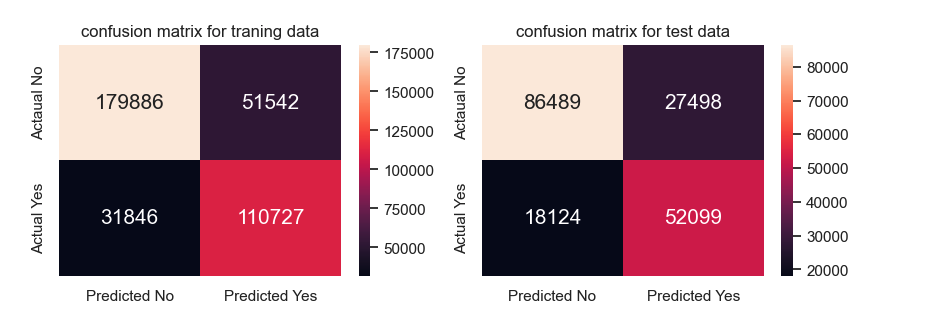


Figure 36 shows the confusion matrix on training data and testing data for the decision tree model just using important features.  
**Figure 36***Confusion Matrix for Decision Trees (using important features)* ****

The results of applying the Random forest model with only the most important features are displayed in Figure 37.  
**Figure 37***Results for Random Forest Model (using important features)*

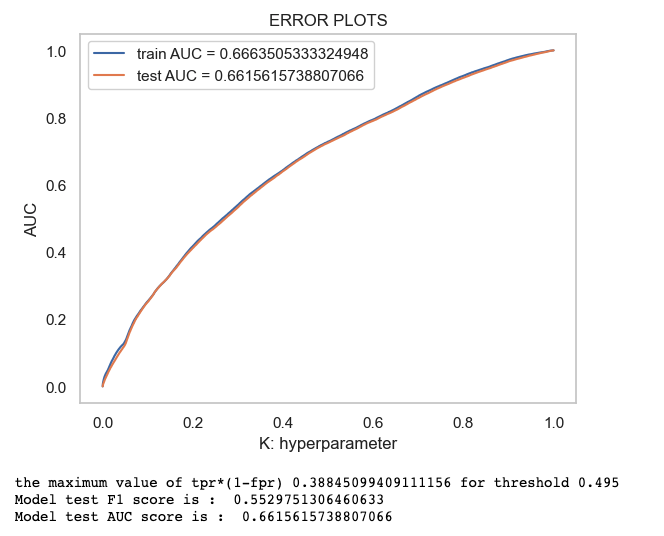
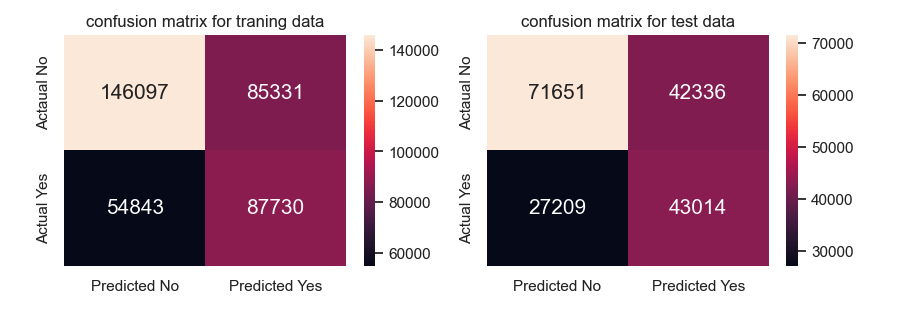
****

Figure 38 shows the confusion matrix on training data and testing data for the Random forest model just using important features.  
**Figure 38***Confusion Matrix for Random Forest Model (using important features)*   


The results of applying the XG-boost model with only the most important features are displayed in Figure 39.  
**Figure 39**  
*Results for XGBOOST Model (using important features)*

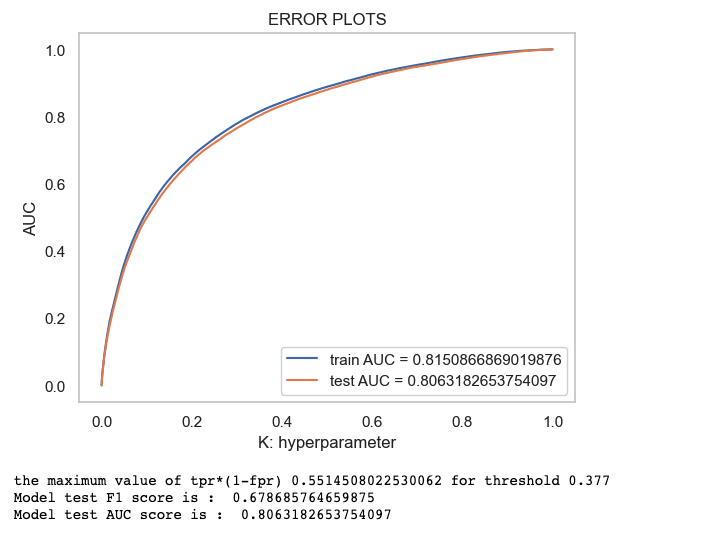
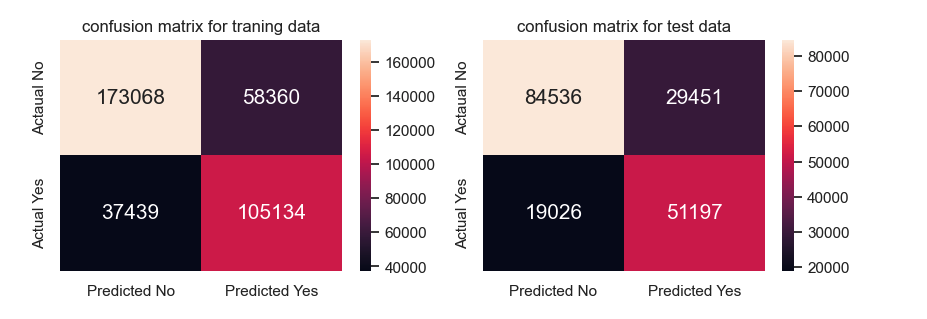


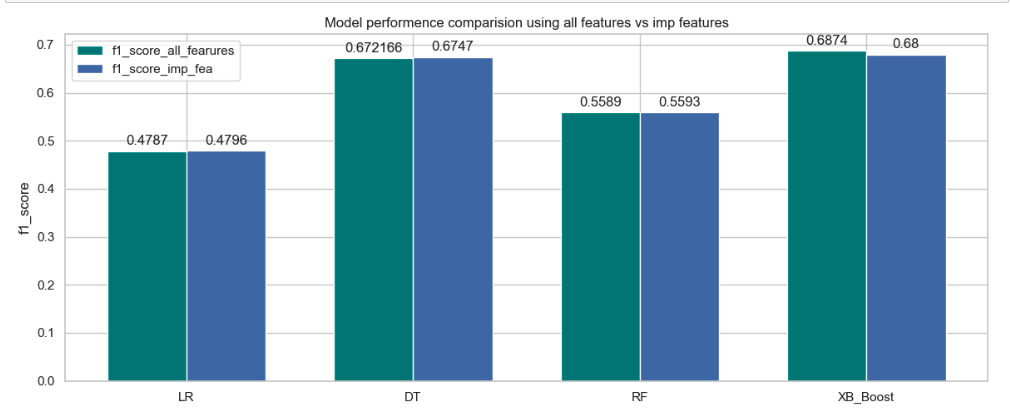
Figure 40 shows the confusion matrix on training data and testing data for the XG-boost model just using important features.

**Figure 40***Confusion Matrix for XG-boost Model (using important features)*   


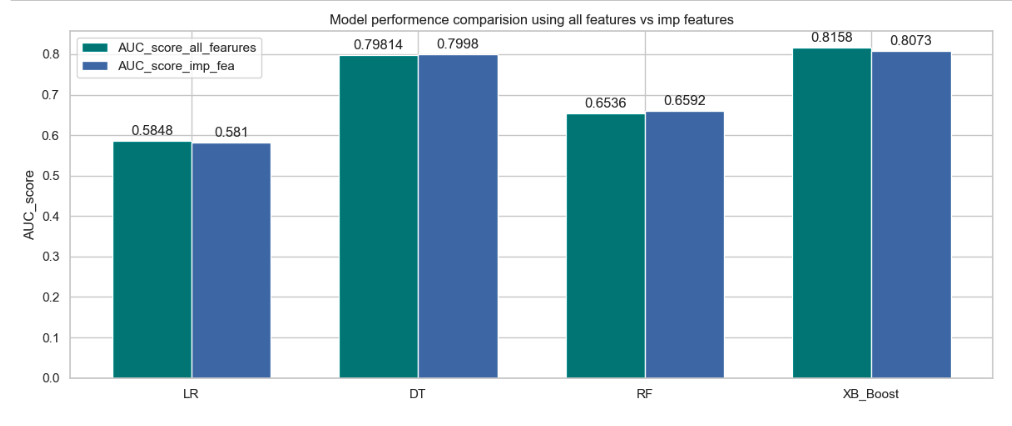
In Table 2 all the models are compared on evaluation metrics such as F1, AUC, and accuracy using just the important features.  
**Table 2***Comparison Using the Important Features*

| **Model** | **Hyperparameter** | **Accuracy on test data** | **F1 on Test** | **AUC** |
| --- | --- | --- | --- | --- |
| Logistic Regression | {'C': 1000.0, 'penalty': 'l2'}. | 0.6287 | 0.48406 | 0.5846 |
| Decision Tree | 'max\_depth' = 50 and 'min\_samples\_split' = 270 | 0.7525 | 0.6954 | 0.8227 |
| Random Forest | - n\_estimators = 500  - max\_features = 'auto'  - max\_depth = 8  - criterion = 'entropy'  - class\_weight = | 0.6352 | 0.5529 | 0.6615 |
| XGBoost | 'n\_estimators': 50, 'eta': 0.3} | 0.7519 | 0.6786 | 0.8063 |

In a comparison of models using all features versus important features, the F1 score is used as the evaluation metric. The F1 score combines precision and recall, providing a balanced measure of a model's performance in binary classification tasks. A higher F1 score indicates better performance in terms of both precision and recall. Below Figure 41 shows the comparison of the model considering the F1 score.  
**Figure 41***Comparison using All Features vs Important Features (F1 score)*



In a comparison of models using all features versus important features, the F1 score is used as the evaluation metric. The F1 score combines precision and recall, providing a balanced measure of a model's performance in binary classification tasks. A higher F1 score indicates better performance in terms of both precision and recall. Below Figure N shows the comparison of the model considering the F1 score. Models are also compared considering AUC\_score. Figure 42 shows the performance of models.  
**Figure 42***Comparison Using All Features vs Important Features(AUC score)*



1. **Deployment**

In the deployment step, our machine learning model moves from the development environment to the production environment, where it is prepared to examine fresh data and produce predictions of property values. To enable retrieval and usage later, we first serialize the learned model using formats like pickle, ONNX, or PMML. The model is then ready for deployment, which can take place on a local machine, dedicated servers, or cloud infrastructure like Google Cloud, AWS. An access point, like an API endpoint, is required for users or systems to communicate with the deployed model and get predictions. Establishing maintenance methods is necessary to guarantee the model's continuous efficacy after deployment. Among these include keeping an eye on performance and periodically updating the model with fresh data. We can increase the effectiveness and performance of the machine learning model in practical settings by putting in place a reliable deployment process. Future work will include creating an interface that is simple to use and straightforward for researchers and healthcare professionals to interact with the system. A simple, user-friendly design with clearly labeled sections and simple controls.

1. **Source Code Link**

The source code can be accessed on GitHub using the following link:

<https://github.com/rasho330/DATA-245-ML>

**Conclusion**

In summary, this project has successfully developed a fraud detection model that accurately predicts potential fraudulent healthcare providers based on their filed claims. By analyzing various features such as diagnosis codes, procedure codes, reimbursement amounts, and beneficiary information, the model demonstrates its effectiveness in detecting and preventing healthcare fraud. This brings significant benefits to insurance companies and governments, including cost reduction and ensuring timely claims processing for legitimate customers. Additionally, the imbalanced nature of the dataset was addressed by implementing evaluation metrics such as precision, recall, and F1 score, along with the binary confusion matrix, to ensure the model's accuracy in identifying potential fraudulent providers while minimizing false positives. Overall, this project contributes to the ongoing fight against healthcare fraud, promoting integrity in the healthcare system and safeguarding the interests of both providers and beneficiaries.

**Future Scope**

There are numerous ways to improve the fraud detection system created in this project in terms of future scope. These include working with law enforcement organizations, developing predictive analytics based on past behavior, integrating new data sources like claims data and electronic health records, implementing real-time detection using machine learning algorithms, and enhancing the user interface and reporting features. Accuracy would increase, proactive fraud protection would be possible, prospective fraud trends would be identified, teamwork would increase efficacy, and user-friendliness and accessibility would be improved. By pursuing these avenues, the initiative can enhance the fraud detection system's capabilities, aiding in the battle against healthcare fraud and fostering system integrity.

**References**

Bauder, R. A., Da Rosa, R., & Khoshgoftaar, T. M. (2018b). Identifying Medicare Provider Fraud with Unsupervised Machine Learning. *Information Reuse and Integration*. https://doi.org/10.1109/iri.2018.00051

Bauder, R. A., & Khoshgoftaar, T. M. (2017). *Medicare Fraud Detection Using Machine Learning Methods*. https://doi.org/10.1109/icmla.2017.00-48

Johnson, J. M., & Khoshgoftaar, T. M. (2022). *Healthcare Provider Summary Data for Fraud Classification*. https://doi.org/10.1109/iri54793.2022.00060

Wu, Y., Wang, Y., Jiang, X., & Chen, Y. (2018). Fraud detection in healthcare: A systematic review. IEEE Access, 6, 8857-8871

*HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS*. (2019, May 9). Kaggle. https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?resource=download&select=Train\_Outpatientdata-1542865627584.csv